OTST: A Two-Phase Framework for Joint Denoising and Remosaicing in RGBW CFA

Zhihao Fan[1,2,*], Xun Wu[1,3,*], Fanqing Meng[4], Yaqi Wu[1,†], Feng Zhang[5,1,†]
[1] Tetras.AI, [2] University of Shanghai for Science and Technology,
[3] Tsinghua University, [4]Tongji University, [5]Shanghai Artificial Intelligence Laboratory.
203590822@st.usst.edu.cn wuxun21@mails.tsinghua.edu.cn mengfanqing33@gmail.com
wuyaqi@tetras.ai zhangfeng@pjlab.org.cn

Abstract

RGBW, a newly emerged type of Color Filter Array (CFA), possesses strong low-light photography capabilities. RGBW CFA shows significant application value when low-light sensitivity is critical, such as in security cameras and smartphones. However, the majority of commercial image signal processors (ISP) are primarily designed for Bayer CFA, research pertaining to RGBW CFA is very rare. To address above limitations, in this study, we propose a two-phase framework named OTST for the RGBW Joint Denoising and Remosaicing (RGBW-JRD) task. For the denoising stage, we propose Omni-dimensional Dynamic Convolution based Half-Shuffle Transformer (ODC-HST) which can fully utilize image's long-range dependencies to dynamically remove the noise. For the remosaicing stage, we propose a Spatial Compressive Transformer (SCT) to efficiently capture both local and global dependencies across spatial and channel dimensions. Experimental results demonstrate that our two-phase RGBW-JRD framework outperforms existing RGBW denoising and remosaicing solutions across a wide range of noise levels. ***In addition, the proposed approach ranks the 2^{nd} place in MIPI 2023 RGBW Joint Remosaic and Denoise competition.***

1. Introduction

Color filter array (CFA) pattern is a critical component placed over the complementary metal-oxide-semiconductor (CMOS) sensor in photographic devices to capture color information. As shown in Figure 1 (b), the Bayer CFA pattern is the most common type of pattern employed in existing photographic equipment due to its simplicity and ability to produce high-quality images. However, smartphone cameras still have notable drawbacks in their ability to deliver professional-grade image quality in stochastic lighting conditions, which affects their overall perceptual quality. To address this dilemma, the RGBW CFA pattern has been proposed. As illustrated in Figure 1 (a), different from Bayer CFA pattern, RGBW CFA pattern consists of red, green, blue and white filters arranged in repeating patterns. Specifically, by capturing all wavelengths of light with the white filter to provide additional brightness information, the RGBW CFA pattern demonstrates a significant improvement in Signal-to-Noise Ratio (SNR) compared to traditional RGB-only sensors, particularly in low-light conditions. Moreover, the inclusion of white pixels does not undermine the integrity of the original RGB color but rather enhances sensitivity by a factor of 1.7. This improvement results in brighter images and reduced color distortion. Notably, several leading Original Equipment Manufacturers (OEMs), such as Transsion, Vivo and Oppo, have recently incorporated RGBW sensors into their flagship smartphones to elevate camera image quality [31].

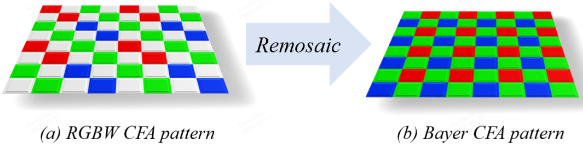


Figure 1. Illustration of two different CFA patterns. (a) RGBW CFA pattern. (b) Bayer CFA pattern. Remosaic processing refers to the conversion of RGBW CFA raw images into Bayer CFA raw images. process illustration.

However, conventional camera ISP pipelines can only work with Bayer raw images, making it impossible to directly utilize the RGBW raw images. Therefore, an interpolation process is necessary to convert the RGBW raw images into Bayer format. The interpolation process referred as remosaicing must achieve two objectives: (1) First, it must obtain the Bayer output with minimal artifacts from

[*] Both authors contributed equally to this research.
[†] Corresponding author.

the RGBW input. (2) Second, it must make full use of the signal-to-noise ratio and resolution advantages of white pixels [31]. Moreover, the remosaicing task becomes more challenging when sensor noise is present, especially in low-light conditions. Recently, various deep learning models have been proposed for the remosaicing task [20, 30, 31]. However, most of them do not consider the issue of noisy data. Since noisy and non-noisy data have different characteristics, the incorporation of both types of data into a network can result in suboptimal denoising outcomes. Therefore, investigating the Joint Denoising and Remosaicing for RGBW CFA pattern (RGBW-JDR) task is important for real-world applications of RGBW CFA pattern.

To address this challenge, we propose a two-phase network named OTST which comprises of a denoising phase and a remosaicing phase. For the denoising phase, we employed an Omni dynamic Convolution [22] to dynamically capture the noise distribution, followed by the Half-Shuffle Transformer (HST) [8] to efficiently capture both local and global spatial-wise similarities and inter-channel correlations. For the remosaicing phase, we propose a Spatial Compressive Transformer (SCT) to efficiently capture both local and global spatial-wise similaries. Our key innovation is the design of a Dual-Spatial MSA (DS-MSA) module that captures local high-frequency details and long-range global dependencies simultaneously. Our proposed framework effectively addresses the challenge of mixed noisy and non-noisy data in a joint manner and outperforms existing methods in both denoising and remosaicing performance. The main contributions of this work are listed as follows:

- We propose a novel two-phase framework to tackle Joint Denoising and Remosaicing for RGBW CFA pattern (RGBW-JDR) task. By this step-by-step manner, our framework outperforms state-of-the-art methods across a wide range of noise levels.
- For remosaicing, we propose a novel Spatial Compressive Transformer (SCT) to capture both local and global spatial dependencies in an efficient manner.
- Our proposed framework achieves the second place in the “RGBW Joint Remosaic and Denoise 2023 @MIPI-challenge” competition.

2. Related Work

2.1. Denoise Raw Images

The denoising of images is a fundamental task in the field of image processing and computer vision. Traditional techniques rely on prior information, such as non-local mean [5], sparse coding [2, 15, 24], BM3D [12], among others. Recently, with the advent of convolutional neural networks, there has been an increased focus on developing end-to-end denoising networks. Advanced network architectures have led to a plethora of CNN-based denoising methods, which are primarily applied in the RGB domain [11, 19, 21]. However, when these methods are directly applied to RAW images, performance suffers because the shape and distribution of noise differ significantly between the RAW and RGB domains. Several public datasets have been proposed for image denoising in the RAW domain [1, 3, 9], and some convolutional neural networks [9, 17, 19] have shown promising results in these datasets. However, obtaining pairs of noise and real images is a laborious task, and thus, generating more realistic RAW domain noise data has become a crucial research topic. Several approaches have been proposed, including Gauss-Poisson distribution noise [16,28], Gaussian mixture models [34], and in-camera process simulation. Wang et al. [28] proposed a method for constructing a Gauss-Poisson noise model and a network to achieve denoising.

2.2. Joint Remosaicing and Denoising

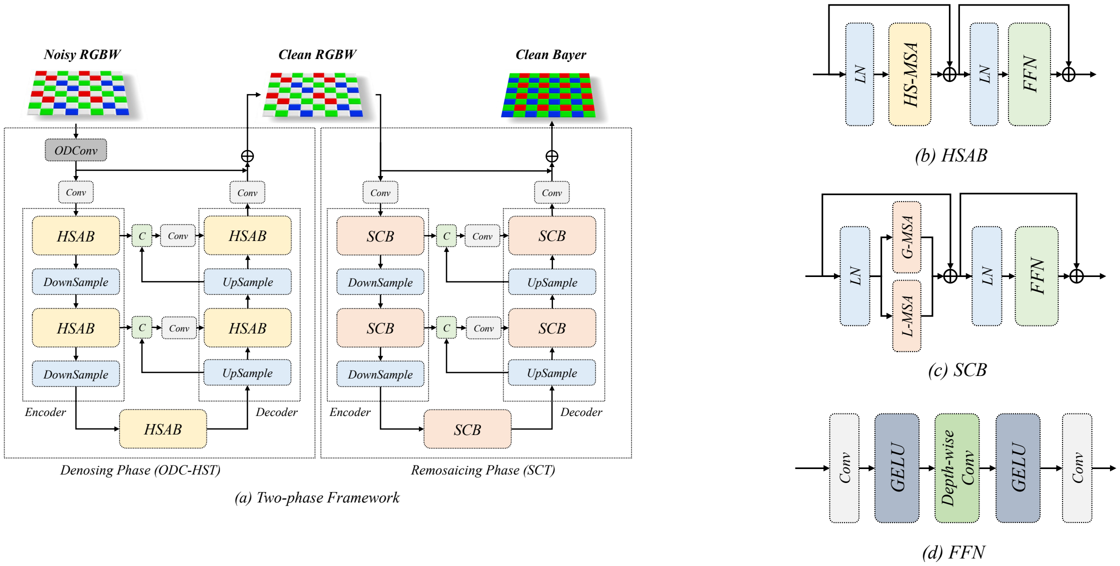
In recent years, denoising and remosaicing have emerged as critical tasks in the field of image processing, attracting significant attention from researchers. Deep learning techniques have enabled researchers to propose various models that achieve state-of-the-art performance in diverse scenarios. Notably, the DRUNet [13] proposed by the “op-summer-Po” team emerged as the top performer in the first “MIPI 2022 Challenge on RGBW Sensor Re-Mosaic” [31]. While this model delivers satisfactory denoising quality and speed, it requires a large amount of high-quality training data to attain optimal performance and may exhibit suboptimal performance in the presence of Poisson noise, which requires specific techniques for elimination to avoid artifacts or distortions.In contrast, the second 2nd team “HIT-IIL” employed the NAFNet [10] model, which can handle various types of noise. However, this model uses a computationally expensive non-local attention mechanism that may be high hardware overhead. The third-place team utilized the Transformer-based Unet network structure and employed the Multi-ResTransformer (MResT) block for each layer of the encoding and decoder instead of the residual convolutional block. However, this method involves a large number of network parameters and takes a long time to train and inference.

3. Method

In this section, we provide a description of the formulation for our two-phase framework OTST. Following this, we present detailed explanations of the key components encompassed within these two phases.

3.1. Two-Phase Framework Formulation

In this work, our framework aims to address the RGBW-JDR task in a step-by-step manner. To achieve this, as



(a) Two-phase Framework

(b) HSAB

(c) SCB

(d) FFN

Figure 2. (a) Illustration of our proposed two-phase RGBW-JDR framework OTST. The full framework consists of two sequential phases, *i.e.*, the denoising phase and remosaicing phase. Each phase contains a U-shaped structure Transformer. (b) HSAB consists of an FFN, an HS-MSA, and two layer normalization. (c) SCB consists of an FFN, two layer normalization, parallel-connected L-MSA and G-MSA. (d) Components of FFN.

shown in Figure 2 (a), we have designed two sequential phases within our OTST. The first phase named denoising phase focuses on eliminating the noise present in the RGBW raw images, while the second phase named remosaicing phase is concerned with converting the cleaned RGBW raw images into Bayer raw images.

In detail, for denoising phase, given a noisy RGBW raw image $\boldsymbol{Y}_{\sigma} \in \mathbb{R}^{H\times W}$, a denoising method noted as $\mathbf{F}_{\gamma}$ is employed to eliminate the noise and further restore the clean RGBW raw image $\boldsymbol{X}_{*} \in \mathbb{R}^{H\times W}$ from $\boldsymbol{Y}_{\sigma}$, *i.e.*,

$$\boldsymbol{X}_{*} = \mathbf{F}_{\gamma}\left(\boldsymbol{X}_{r}\right). \qquad (1)$$

After that, $\boldsymbol{X}_{*}$ is took as the input of remosaicing phase and a remosaicing method $\mathbf{F}_{\odot}$ is adopted to reconstruct the corresponding Bayer CFA from it. The ensuing expression is given as follows:

$$\boldsymbol{X} = \mathbf{F}_{\odot}\left(\boldsymbol{X}_{*}\right). \qquad (2)$$

Finally, we get the clean Bayer CFA raw image $\boldsymbol{X} \in \mathbb{R}^{H\times W}$. The whole two-phase framework can be formulated as:

$$\boldsymbol{X} = \mathbf{F}_{\odot}\Bigg(\mathbf{F}_{\gamma}\Big(\boldsymbol{Y} + \mathcal{N}\left(0,\ \boldsymbol{Y}\cdot\sigma_s^2+\sigma_c^2\right)|\boldsymbol{\theta}_{\gamma}\Big)|\boldsymbol{\theta}_{\odot}\Bigg). \qquad (3)$$

Here $\boldsymbol{\theta}_{\odot}$, $\boldsymbol{\theta}_{\gamma}$ denote the learnable parameters in $\mathbf{F}_{\odot}$ and $\mathbf{F}_{\gamma}$. By the designed step-by-step manner, our OTST achieves outstanding performance.

3.2. Denoising Phase (ODC-HST)

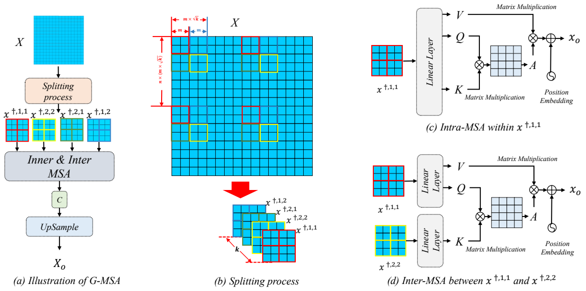
We propose ODC-HST to play the role of $\mathbf{F}_{\gamma}$, which consists of two sequential modules: an Omni-dimensional Dynamic Convolution (ODC) [22] to obtain the noise distribution of the entire raw image, and a Half-Shffle Transformer (HST) [8] to eliminates the noise. In this section, we present the detailed structures of these two modules in Section 3.2.1 and Section 3.2.2.

3.2.1 Omni-dimensional Dynamic Convolution (ODC)

The initial module plays an important role on low-level recovery task and can significant influence the reconstruction quality. Nonetheless, most existing denoising methods typically employ traditional convolution to extract initial features, which exhibits limited representation power due to its static computational manner.

To address this problem, we utilize Omni-dimensional Dynamic Convolution (ODC) [22] to play the role of the initial module in our framework. By leveraging the multi-dimensional attentions along all four dimensions of the kernel space present in ODC, our framework achieves a more generalized dynamic convolution and obtains stronger denoising capabilities.

In detail, taking the noisy RGBW raw image $\boldsymbol{Y}_{\sigma} \in \mathbb{R}^{H\times W}$ as input, for i-th convolutional kernel $\boldsymbol{W}_i \in \mathbb{R}^{k\times k}$, ODC predicts four types of attention scalars for it, *i.e.*, kernel-wise scalar $\alpha_{wi} \in \mathbb{R}$, spatial-wise scalar $\alpha_{si} \in \mathbb{R}^{k\times k}$, input channel-wise scalar $\alpha_{ci} \in \mathbb{R}$ and output channel-wise scalar $\alpha_{fi} \in \mathbb{R}^{\lambda}$. Mathematically, the gen-


X
Splitting process
$x^{\dagger,1,1}$ $x^{\dagger,2,2}$ $x^{\dagger,2,1}$ $x^{\dagger,1,2}$
Inner & Inter MSA
C
UpSample
X_o
$m \times \sqrt{k}$
m
m
$n \times (m \times \sqrt{k})$
X
$x^{\dagger,1,2}$ $x^{\dagger,2,1}$ $x^{\dagger,2,2}$ $x^{\dagger,1,1}$
k
Linear Layer
V
Q
K
Matrix Multiplication
A
x_o
Position Embedding
$x^{\dagger,1,1}$
$x^{\dagger,2,2}$

(a) Illustration of G-MSA

(b) Splitting process

(c) Intra-MSA within $x^{\dagger,1,1}$

(d) Inter-MSA between $x^{\dagger,1,1}$ and $x^{\dagger,2,2}$

Figure 3. (a) Visual illustration of G-MSA (b) Splitting process: G-MSA first samples k visual tokens from input $\boldsymbol{X}$ with a dilation rate of $n = 2$. Tokens with the same color borders belong to the same partition $\boldsymbol{x}^{\dagger,i,j}$. (c) Illustration of intra-MSA for $\boldsymbol{x}^{\dagger,1,1}$. (d) Illustration of inter-MSA between $\boldsymbol{x}^{\dagger,1,1}$ and $\boldsymbol{x}^{\dagger,2,2}$.

erated dynamic kernel $\boldsymbol{W}_i^{\dagger}$ can be formulated as:

$$\boldsymbol{W}_i^{\dagger} = \alpha_{wi} \odot \alpha_{fi} \odot \alpha_{ci} \odot \alpha_{si} \odot \boldsymbol{W}_i \tag{4}$$

Then the full dynamic convolution operations can be defined as:

$$\boldsymbol{X}_r = \left(\sum_{i=1}^{N} \boldsymbol{W}_i^{\dagger} \right) \otimes \boldsymbol{Y}_{\sigma}. \tag{5}$$

where N denotes the number of kernels contained in ODC. By employing attention mechanisms across the four dimensions of the kernel space, we are able to dynamically model the distribution of noise and consequently acquire superior features.

3.2.2 Half-Shuffle Transformer (HST)

Existing CNN-based denoising methods have achieved impressive results [4, 6, 28]. However, these approaches demonstrate deficiencies in modeling long-range dependencies and non-local similarities. In contrast, Transformer has shown exceptional ability over the past few years in modeling long-range dependencies with great efficacy [14,23,27]. However, applying Transformer directly to denoising tasks may pose two challenges: limited receptive fields and significant computational costs.

To address this problem, we introduce Half-Shuffle Transformer (HST) [8]. As shown in Figure 2 (a) and (b), HST consists of a three-level U-shaped structure constructed using the Half-Shuffle Attention Block (HSAB) as its basic unit. Based on the Half-Shuffle Multi-head Self-Attention (HS-MSA) which contains two parallel branches, *i.e.*, local Branch and non-local Branch, HST combines the advantages of global MSA [14] and local window-based MSA [23] in an efficient manner.

In detail, the input rough noisy feature $\boldsymbol{X}_r \in \mathbb{R}^{H \times W \times \lambda}$ is first projected into query $\mathbf{Q} \in \mathbb{R}^{H \times W \times \lambda}$, key $\mathbf{K} \in \mathbb{R}^{H \times W \times \lambda}$, and value $\mathbf{V} \in \mathbb{R}^{H \times W \times \lambda}$. After that, these three elements are splitted along channel dimension, *i.e.*,

$$\mathbf{Q} = [\mathbf{Q}_L, \mathbf{Q}_N],\ \mathbf{K} = [\mathbf{K}_L, \mathbf{K}_N],\ \mathbf{V} = [\mathbf{V}_L, \mathbf{V}_N], \tag{6}$$

Here $\mathbf{Q}_L, \mathbf{K}_L, \mathbf{V}_L \in \mathbb{R}^{H \times W \times \frac{\lambda}{2}}$ are fed into the local branch to capture local contents, while $\mathbf{Q}_N, \mathbf{K}_N, \mathbf{V}_N \in \mathbb{R}^{H \times W \times \frac{\lambda}{2}}$ pass through the non-local branch to model non-local dependencies.

Local Branch. Inspired by [23], the local branch utilizes non-overlapping shifted $k \times k$ window to split spatial patches $\mathbf{Q}_L^{\dagger}, \mathbf{K}_L^{\dagger}, \mathbf{V}_L^{\dagger} \in \mathbb{R}^{\frac{HW}{k^2} \times k^2 \times \frac{\lambda}{2}}$, and then employs typical channel-wise MSA [7] within these patches to capture local high-frequency details.

Non-local Branch. Inspired by ShuffleNet [33], the non-local branch employs shuffling operation to capture cross-window interactions. Specifically, non-local branch utilizes non-overlapping shifted local $k \times k$ window to capture spatial patches $\mathbf{Q}_N^{\dagger}, \mathbf{K}_N^{\dagger}, \mathbf{V}_N^{\dagger} \in \mathbb{R}^{\frac{HW}{k^2} \times k^2 \times \frac{\lambda}{2}}$. Then their shapes are transformed to $\mathbf{Q}_N^{\ddagger}, \mathbf{K}_N^{\ddagger}, \mathbf{V}_N^{\ddagger} \in \mathbb{R}^{k^2 \times \frac{HW}{k^2} \times \frac{\lambda}{2}}$. By this way, the positions of tokens are shuffled and inter-window dependencies are established. Finally, non-local branch computes typical channel-wise MSA [7] within these transformed patches.

By fusing the outputs from the local and non-local branches, HS-MSA achieves both local and global feature integration with reducing computational cost, thus enhancing the modeling abilities in a cost-effective manner.

3.3. Remosaicing Phase (SCT)

After obtaining the clean RGBW raw image $\boldsymbol{X}_1$, we employ the remosaicing phase $\mathbf{F}_{\odot}$ to convert the RGBW raw image to the Bayer pattern. However, similar to denoising methods, previous remosaicing methods exhibit limitations in capturing long-range dependencies and non-local similarities. Besides, these methods also have some limitations in modeling inter-channel corrections: (1) First, the computational cost of typical channel-wise MSA [7] is relatively high. (2) Second, traditional channel attention (*e.g.*, SENet [18]) has been proved losing high-frequency details [26].

To address above problems, we propose the Spatial Compressive Transformer (SCT), which aims to efficiently model both local-global spatial self-similarities and inter-channel correlation. As shown in Figure 2 (c), to achieve this, the basic unit of SCT, named Spatial Compressive Block (SCB), a Local-Global Dual Spatial-wise MSA (DS-MSA) module to capture both local high-frequency details and long-range global dependencies at the same time. In this section, we present the detailed structures of proposed DS-MSA in Section 3.3.1.

3.3.1 Dual Spatial-wise MSA (DS-MSA)

Our Dual Spatial-wise MSA (DS-MSA) consists of two parallel MSA branches, *i.e.*, Local-wise spatial MSA branch (L-MSA) and Global-wise spatial MSA branch (G-MSA), which can efficiently capture both local and non-local dependencies at the same time.

Local-wise spatial MSA branch. Similar to the local branch presented in Section 3.2.2, we adopt shifted local windows self-attention proposed in [23] to capture finegained local details. Specifically, as shown in Figure 3 (a), by using $k \times k$ local window to limit spatial attention computation in $m \times m$ local patches, L-MSA achieves capturing local high-frequency details with linear computational complexity $\boldsymbol{\Omega}_L = 4NC^2 + 2Nm^2C$ where $N = \frac{H \times W}{k^2}$. $\boldsymbol{\Omega}_L$ is linear when m is fixed, which means the computational complexity of L-MSA is scalable and affordable.

Global-wise spatial MSA branch. To capture global spatial dependencies, there are two methods: (1) The standard spatial-wise MSA [14] has superior ability to build long-range dependencies but suffers from high computational costs on high-resolution images. (2) Pooling operations [25] are usually utilized as efficient methods on the spatial dimension to capture global features. However, pooling operations are not suitable for the demosaicing problem due to the loss of high-frequency information [26]. So, one problem is how to capture global spatial details while saving computational costs.

To achieve this, we propose an efficient Global-wise spatial MSA (simplified as G-MSA). The overall processes of G-MSA is shown in Figure 3 (a). Generally speaking, G-MSA first splits input features into several dilated partitions along spatial dimension. As shown in Figure 3 (b), it is worth nothing that pixels in each partition are not from a local region but subsampled from the whole input feature with a dilation rate. Then G-MSA computes both intra-MSA and inter-MSA between each pairs of partitions to capture global dependencies. Figure 3 (c) $\sim$ (d) provide examples by computing intra-MSA and inter-MSA, respectively. After that, G-MSA concatenates these outputs along channel dimension. Finally, a upsample×2 module is employed to scale up the acquired features to match the spatial dimensionality of the original input.

By aggregating the outputs of the L-MSA and G-MSA branches, our DS-MSA achieves the ability to efficiently capture both local and global spatial information simultaneously.

3.4. Loss Function

In many image reconstruction and enhancement tasks, the mean absolute error (MAE) loss is widely used. In this work, we denoise and remosaic the loss function using MAE loss:

$$\begin{cases} \mathcal{L}_D = \|\boldsymbol{X}_* - \boldsymbol{X}_0\|_1 \\ \mathcal{L}_R = \|\boldsymbol{X} - \boldsymbol{I}_{gt}\|_1 + \lambda\|\boldsymbol{X}_{rgb} - \boldsymbol{I}_{rgb}\|_1 \end{cases} \tag{7}$$

Where $\boldsymbol{X}_*$ and $\boldsymbol{X}_0$ denotes the clean output of denose phase and 0dB RGBW. $\boldsymbol{X}$ and $\boldsymbol{I}_{gt}$ represent the reconstructed Bayer of remosaic model and ground truth Bayer respectively. $\boldsymbol{X}_{rgb}$ and $\boldsymbol{I}_{rgb}$ indicates $\boldsymbol{X}$ and $\boldsymbol{I}_{gt}$ after the official ISP to obtain RGB images. λ is a hyper-parameter tuning $\mathcal{L}_R$.

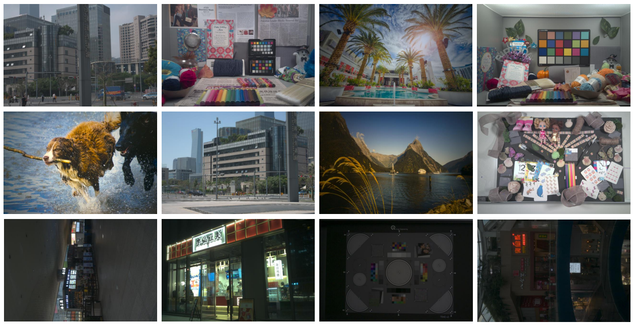


Figure 4. Visualization of MIPI 2023 Challenge on RGBW Joint Remosaic and Denoise dataset. This dataset contains multiple scenes (i.e., natural scene, dark scene).

4. Experiment

4.1. Datasets

In this study, we evaluated our proposed framework using a dataset of high-quality RGBW and Bayer image pairs provided by the "MIPI 2023-RGBW Joint Remosaic and Denoise" competition. The dataset consists of 100 scenarios, of which 70 were used for training, 15 for validation, and the remaining 15 for testing. All images in the dataset have a fixed resolution of 1200×1800 pixels, and each RGBW data piece contains three different noise levels: 0dB, 24dB, and 42dB. Sample images from various scenarios are depicted in Figure 4.

4.2. Evaluation Metrics

The evaluation of our algorithm was conducted in two parts: firstly, the comparison of the restored Bayer image with the ground truth Bayer image, and secondly, the comparison of the RGB results generated from the Bayer image using a simple ISP. To measure the former, we employed the KLD as the evaluation metrics, while for the latter, we utilized the PSNR, SSIM [29], and LIPIS [32] as evaluation metrics. To provide an intuitive measure of the effectiveness of our algorithm and overall image quality, we follow the approach of MIPI challenge and used M4 scores as the comprehensive evaluation metrics. The calculation method of M_4 is shown in Equation 8.

$$M_4 = PSNR \times SSIM \times 2^{1-LPIPS-KLD} \tag{8}$$

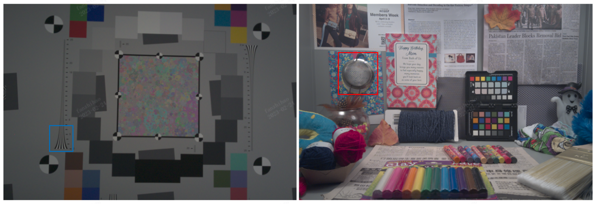


Figure 5. Valid the images visualized in the set, with 075 on the left and 076 on the right

Table 1. The final testing results of MIPI 2023 Challenge on RGBW Joint Remosaic and Denoise. The maximum value is in bold and our results are highlighted in gray.

Rank	Team	PSNR↑	SSIM↑	LPIPS↓	KLD↓	M4↑
1	ChongQB	38.55	**0.98**	**0.07**	**0.06**	**69.047**
2	fanzhihao(Ours)	**38.724**	0.9739	0.0804	0.0662	68.139
3	CD_luo	38	0.96	0.07	0.07	66.213
4	Legless_bird	37.78	0.97	0.09	0.07	65.599
5	blindbox	37.71	0.96	0.08	0.07	65.253
6	jiangchengzhi	36.69	0.96	0.09	0.07	63.050
7	Senta	36.25	0.97	0.12	0.07	61.647
8	VLAB	35.78	0.96	0.12	0.07	60.221
9	xiajiyuan	35.8	0.96	0.13	0.07	59.838
10	gymlab	35.13	0.96	0.13	0.07	58.718

The M_4 evaluation metrics ranges from 0 to 100, with a higher scores indicating a higher image quality.

4.3. Implementation Details

The training details are presented as follows: the model is implemented in Pytorch and performed on 8 Titan XP graphical processing units (GPUs). The model is optimized using an Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, learning rate = $1e-4$ with a batch size of 4 and a patch size of 128. To substantiate the efficacy of our two-phase approach, we first compare the results of processing all data directly. Then we evaluate the pertinent benchmark models in comparison to our proposed two-phase model.

In this paper, we compared our proposed model with the OPPO team in the “MIPI 2022 Challenge on RGBW Sensor Re-mosaic” [31] challenge and some other state-of-the-art methods in other fields. Specifically, we benchmarked our proposed model against the MSTP [7] which is the winning solution in the NTIRE-CVPR hyperspectral reconstruction competition and the HST [8] which has shown exceptional performance in spectral compression. To ensure a fair comparison, all models are optimized and evaluated in the same training strategy.

4.4. Experimental Results

4.4.1 Testing Results of MIPI 2023 Challenge on RGBW Joint Remosaic and Denoise

The proposed two-phase framework OTST ranked 2^{nd} in “MIPI 2023 Challenge on RGBW Joint Remosaic and Denoise”. The final comparison results on testing set are summarized in Table 1. It is worth nothing that our OTST achieves the best performance in term of PSNR. Besides, Our OTST exhibits very similar performance to the first-place method in term of SSIM and KLD. These results demonstrate that our OTST can be considered as a commendable solution for RGBW-JDR task.

Table 2. Ablation study for RGBW-JDR approaches. Our two-phase framework OTST outperforms in most cases the four end-to-end competitors SCT, HST, MSTP, OPPO, where SCT is the model we designed for remosaic, and we highlight our approach in gray.

	σ	HST	MSTP	OPPO	SCT	OTST
PSNR↑	0	37.299	39.125	36.975	39.464	**41.897**
	24	34.841	34.547	34.353	35.648	**36.727**
	42	31.445	31.357	30.805	31.638	**32.440**
	AVG	34.528	35.010	34.044	35.583	**37.022**
SSIM↑	0	0.9800	0.9822	0.9768	0.9824	**0.9852**
	24	0.9606	0.9588	0.9566	0.9619	**0.9650**
	42	0.9248	**0.9280**	0.9167	0.9272	0.9273
	AVG	0.9551	0.9564	0.9500	0.9572	**0.9591**
LPIPS↓	0	0.0403	0.0290	0.0458	0.0267	**0.0211**
	24	0.1328	0.1342	0.1379	**0.1217**	0.1221
	42	0.2091	0.2120	0.2367	0.2206	**0.1731**
	AVG	0.1274	0.1251	0.1401	0.1230	**0.1054**
KLD↓	0	**0.0237**	0.0412	0.0587	0.0582	0.0858
	24	**0.0289**	0.0328	0.0533	0.0325	0.0451
	42	0.0336	0.0591	0.0698	**0.0329**	0.0552
	AVG	**0.0287**	0.0444	0.0606	0.0412	0.0620
M4↑		59.7007	60.2490	56.9413	61.4066	**63.8079**

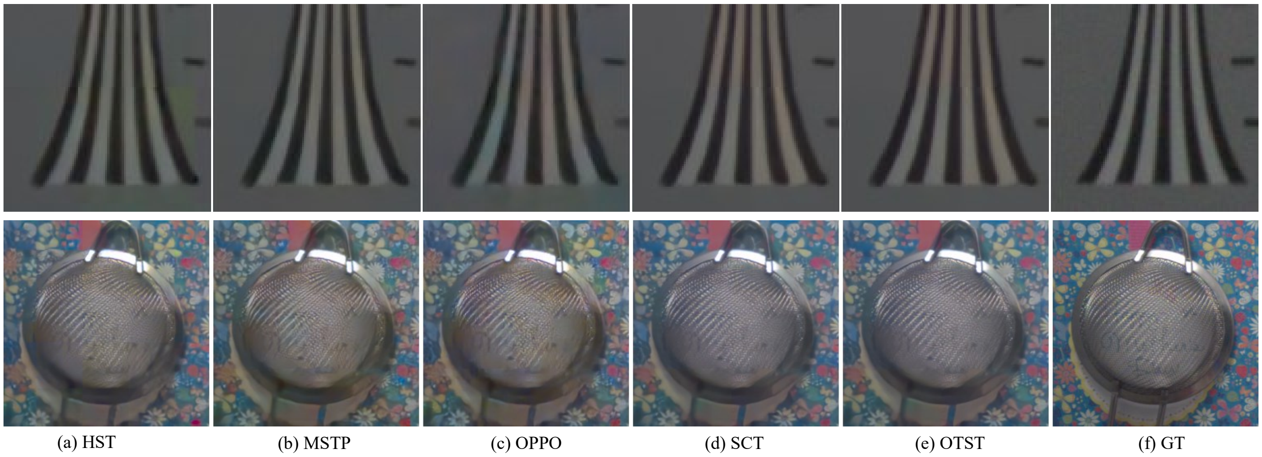
Table 3. Ablation study for denoising performance comparison between four methods: ODC-HST, MSTP, HST and SCT. The maximum value is shown in bold and the results of our denoising method are highlighted in gray.

	PSNR↑	
σ	24	42
SCT	42.8605	37.2287
HST	42.9075	37.3695
MSTP	42.3355	36.858
ODC-HST	**43.2459**	**37.8719**

4.4.2 Ablation Study

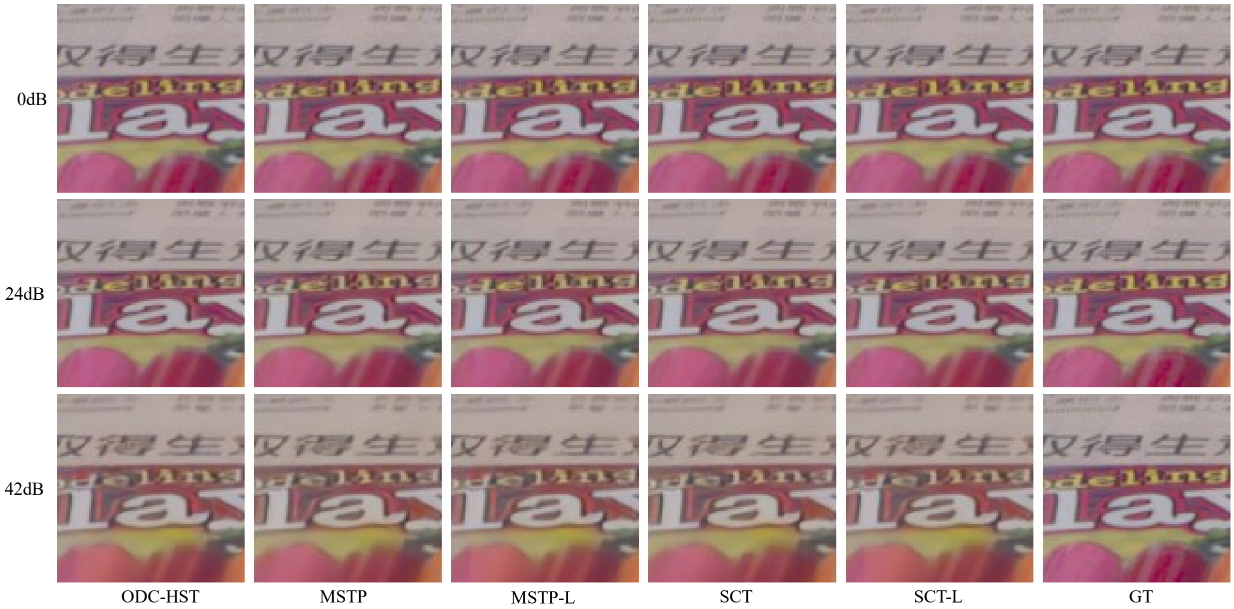
In this section, we perform ablation studies to verify the effectiveness of our proposed two-phase manner and main components in OTST.

Two-phase Manner To verify the effectiveness of our proposed two-stage approach, we compared it with four other end-to-end RGBW-JDR methods (i.e. SCT, HST, MSTP and OPPO), while HST and SCT are denoising and remosaicing stages in OTST respectively. The corresponding results shown in Table 2 indicate that from the comparison of the results for SCT, HST, MSTP and OPPO we can



(a) HST (b) MSTP (c) OPPO (d) SCT (e) OTST (f) GT

Figure 6. Comparison of images recovered at 42dB. The first row is the local picture in the red box in the image 074, and the second row is the local picture in the blue box in the image 075. From left to right, we represent (a) SCT, (b) HST, (c) MSTP, (d) OPPO, (e) OTST, (f) GT.



ODC-HST MSTP MSTP-L SCT SCT-L GT

Figure 7. The figure shows a partial enlarged view of the position of the color pen in 075. the first line is the comparison of 0dB, the second line is the comparison of 24dB, and the third line is the comparison of 42dB. The left-to-right methods of each row are ODC-HST, MSTP, MSTP-L, SCT, SCT-L, GT.

find that our designed SCT has better performance in end-to-end. In addition to this, the OTST in the table represents our proposed two-stage approach, which can be seen to outperform other end-to-end approaches in most settings. Furthermore, as shown in Figure 6, we provide a quantitative comparison to demonstrate the effectiveness of our two-phase approach. We observe that our OTST can recover better high-frequency information (such as texture details) and fine-grained structural information than the other competitors. Both qualitative and quantitative comparison results demonstrate the effectiveness of our proposed two-phase manner.

Denoising. To validate the effectiveness of our denoising method ODC-HST, we compare it with several other models, including HST [8], MSTP [7], and SCT. Table 3 shows that our ODC-HST model achieves best denoising performance in term of PSNR scores, demonstrating the effectiveness of our designed denoising module ODC-HST.

Remosaicing. We compare the remosaicing performance of our proposed SCT with other different models, including ODC-HST and MSTP, while MSTP-L and SCT-L refer to the versions of MSTP and SCT with larger parameters.

Table 4. Ablation study for remosaicing performance comparison between five methods: ODC-HST, MSTP, MSTP-L, SCT and SCT-L. The maximum value is shown in bold, and the results of our remosaicing method are highlighted in gray.

	σ	ODC-HST	MSTP	MSTP-L	SCT	SCT-L
PSNR↑	0	40.373	41.118	41.696	41.897	**42.120**
	24	36.211	36.537	36.619	**36.727**	36.702
	42	32.231	32.421	32.337	**32.440**	32.343
	AVG	36.272	36.692	36.884	37.022	**37.053**
SSIM↑	0	0.984	**0.985**	**0.985**	**0.985**	**0.985**
	24	0.964	**0.965**	**0.965**	**0.965**	**0.965**
	42	0.926	**0.928**	0.927	0.927	0.927
	AVG	0.958	**0.959**	**0.959**	**0.959**	**0.959**
LPIPS↓	0	0.025	0.023	0.022	**0.021**	**0.021**
	24	0.125	0.123	0.123	**0.122**	0.123
	42	0.177	0.174	0.176	**0.173**	0.176
	AVG	0.109	0.107	0.107	**0.105**	0.107
KLD↓	0	**0.020**	0.047	0.065	0.086	0.048
	24	**0.032**	0.037	0.039	0.045	0.037
	42	**0.039**	0.050	0.053	0.055	0.049
	AVG	**0.030**	0.045	0.052	0.062	0.044
M4↑		63.749	63.984	63.950	63.808	**64.644**

Table 5. Ablation study of dilation rate n in G-MSA for remosaicing performance. Without loss of generality, we keep local-window size $m = 8$ in L-MSA. The MACs is computed only for G-MSA module. The maximum value is in bold.

n	MACs	PSNR↑	SSIM↑	LPIPS↓	KLD↓	M4↑
1	844.28M	**37.187**	**0.966**	**0.086**	**0.051**	**65.330**
2	211.13M	37.022	0.959	0.105	0.062	63.808
4	**52.79M**	36.848	0.945	0.111	0.073	61.304

To ensure a fair comparison, we employ two-phase manner for all the remosaicing methods and select ODC-HST as denoising module due to it's strong denoising abilities. Among the evaluated methods, our SCT achieves the best performance across a wide range of noise levels. Furthermore, we also offer qualitative comparison at the color pen below number 075 in the valid datasets (as shown in Figure 7). The diagram reveals that our SCT slightly outperforms the others in letter position at 24 dB and 42 dB. Both qualitative and quantitative comparison results demonstrate the effectiveness of our proposed remosaicing method SCT. Besides, we also validate the effectiveness of proposed G-MSA for capturing long-range dependencies. As we can see in Table 5, by adopting specifically designed feature subsampling mode and cross-scale feature conversion to obtain whole information, our G-MSA achieves better efficiency (75% and 94% computation cost reduction for $n = 2$ and 4, respectively) with a slightly performance drop when dilation rate n increasing.

4.5. Ensemble

As discussed above, each individual model demonstrates promising performance. To generate more robust and accurate results, we introduce ensemble learning. Ensemble learning is a powerful and flexible approach that can improve the performance and reliability of prediction models in a wide range of applications

In our final submission, similar to [7], we adopt two different ensemble strategies. Our approach consists of two main ensemble strategies. First, we utilize a multi-scale ensemble, which involves training the same type of model with different patch sizes (e.g., SCT and SCT-L) and then averaging the outputs to enhance the restoration quality. Second, we employ a top-k multi-model ensemble, which involves training different types of models (e.g., SCT and MSTP) and then averaging their outputs. These ensemble strategies offer several advantages. The multi-scale ensemble can improve the robustness of the model by incorporating multiple patch sizes and leveraging their respective strengths. The top-k multi-model ensemble can enhance the diversity and generalizability of the model by combining different types of models and generating a more comprehensive representation of the underlying patterns in the data. Together, these strategies provide a powerful and flexible approach for improving the performance and reliability of restoration models.

In the final test phase, based on above two ensemble strategies, our approach achieved a marked improvement from M4: 65.913 to 68.139, which reflects the effectiveness of our ensemble strategies.

5. Conclusion

In this paper, we introduce a two-phase framework aiming to solve Joint Denoisng and Remosaicing for the RGBW CFA pattern (RGBW-JDR) task, which consists of a denoising phase and a remosaicing phase. For the denoising phase, we utilize a Half-Shuffle Transformer with an Omni dynamic Convolution to accurately capture the noise distribution and then eliminate it. For remosaicing phase, we propose a Spatial Compressive Transformer (SCT) to efficiently capture both local and global dependencies across spatial and channel dimensions. Experimental results demonstrate that our two-phase RGBW-JDR framework OTST significantly outperforms existing RGBW denoising and remosaicing solutions across a wide range of noise levels. These results highlight the effectiveness of our approach in producing high-quality images with reduced noise levels, making it a valuable tool in the field of image processing.

Acknowledgements: This work is partially supported by the National Key RD Program of China (NO.2022ZD0160100).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 2

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006. 2

[3] Josue Anaya and Adrian Barbu. Renoir–a dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, 51:144–154, 2018. 2

[4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 4

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 2

[6] Jaeseok Byun, Sungmin Cha, and Taesup Moon. Fbi-denoiser: Fast blind image denoiser for poisson-gaussian noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5768–5777, 2021. 4

[7] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022. 4, 6, 7, 8

[8] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *arXiv preprint arXiv:2205.10102*, 2022. 2, 3, 4, 6, 7

[9] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2

[10] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 2

[11] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 2

[12] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image restoration by sparse 3d transform-domain collaborative filtering. In *Image Processing: Algorithms and Systems VI*, volume 6812, pages 62–73. SPIE, 2008. 2

[13] Sripad Krishna Devalla, Prajwal K Renukanand, Bharathwaj K Sreedhar, Giridhar Subramanian, Liang Zhang, Shamira Perera, Jean-Martial Mari, Khai Sing Chin, Tin A Tun, Nicholas G Strouthidis, et al. Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomedical optics express*, 9(7):3244–3265, 2018. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5

[15] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006. 2

[16] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 2

[17] Keigo Hirakawa and Thomas W Parks. Joint demosaicing and denoising. *IEEE Transactions on Image Processing*, 15(8):2146–2157, 2006. 2

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. pages 7132–7141, 2018. 4

[19] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. *Advances in neural information processing systems*, 21, 2008. 2

[20] Younghoon Kim, Jungmin Lee, SungSu Kim, Jiyun Bang, Dagyum Hong, TaeHyung Kim, and JoonSeo Yim. Camera image quality tradeoff processing of image sensor re-mosaic using deep neural network. *Electronic Imaging*, 2021(9):206–1, 2021. 2

[21] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 2

[22] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. 2, 3

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 5

[24] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *2009 IEEE 12th international conference on computer vision*, pages 2272–2279. IEEE, 2009. 2

[25] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *arXiv preprint arXiv:2205.13213*, 2022. 5

[26] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021. 4, 5

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 4

[28] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 2, 4

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[30] Xun Wu, Zhihao Fan, Jiesi Zheng, Yaqi Wu, and Feng Zhang. Learning to joint remosaic and denoise in quad bayer cfa via universal multi-scale channel attention network. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 147–160. Springer, 2023. 2

[31] Qingyu Yang, Guang Yang, Jun Jiang, Chongyi Li, Ruicheng Feng, Shangchen Zhou, Wenxiu Sun, Qingpeng Zhu, Chen Change Loy, Jinwei Gu, et al. Mipi 2022 challenge on rgbw sensor re-mosaic: Dataset and report. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 36–45. Springer, 2023. 1, 2, 6

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[33] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 4

[34] Fengyuan Zhu, Guangyong Chen, and Pheng-Ann Heng. From noise modeling to blind image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2016. 2